

Geometria Proiettiva della Traduzione e degli Spazi Interlinguistici

Introduzione al Paradigma Geometrico nella Traduzione

La traduzione automatica contemporanea ha subito una trasformazione radicale negli ultimi decenni, passando da approcci puramente simbolici e basati su regole a metodologie fondate sulla rappresentazione geometrica del significato linguistico. Questa rivoluzione concettuale trova le sue radici nell'intuizione fondamentale che il linguaggio naturale, nelle sue molteplici manifestazioni idiomatiche, può essere modellato attraverso strutture geometriche immerse in spazi vettoriali ad alta dimensionalità. La geometria proiettiva della traduzione emerge come framework teorico unificante che consente di interpretare i fenomeni interlinguistici attraverso le lenti della matematica differenziale, della topologia algebrica e della teoria delle varietà.

Il concetto centrale che governa questa prospettiva è quello di spazio di embedding, ovvero una rappresentazione continua del lessico e delle strutture sintattiche di una lingua all'interno di uno spazio vettoriale euclideo di dimensione finita. Questi spazi, tipicamente caratterizzati da dimensionalità comprese tra 100 e 1024 dimensioni, costituiscono il substrato geometrico sul quale si articolano le operazioni di traduzione. La scelta della dimensionalità rappresenta un compromesso delicato tra espressività rappresentazionale e trattabilità computazionale, dove dimensioni più elevate permettono di catturare sfumature semantiche più sottili a costo di una maggiore complessità negli algoritmi di apprendimento.

Spazi Vettoriali del Linguaggio e Embedding Monolingue

Prima di affrontare la complessità degli spazi multilingue, è necessario comprendere la struttura geometrica degli embedding monolingue. La tecnica seminale Word2Vec, introdotta da Mikolov e collaboratori nel 2013, ha rivoluzionato la rappresentazione distribuita delle parole attraverso due architetture complementari: il modello Continuous Bag-of-Words (CBOW) e il modello Skip-gram. Entrambi gli approcci si fondano sull'ipotesi distribuzionale di Harris, secondo cui parole che occorrono in contesti simili tendono ad avere significati correlati.

Nel modello Skip-gram, l'obiettivo di apprendimento consiste nel massimizzare la probabilità condizionata delle parole di contesto data una parola target. Formalmente, dato un corpus di parole $w_1, w_2, \dots, w_T$, l'obiettivo è massimizzare la funzione di log-verosimiglianza media:

$$L = (1/T) \sum_{t=1}^T \sum_{(-c \leq j \leq c, j \neq 0)} \log p(w_{t+j} | w_t)$$

dove c rappresenta la dimensione della finestra di contesto. La probabilità condizionata viene parametrizzata attraverso la funzione softmax:

$$p(w_o | w_i) = \exp(v_o^T \cdot v_i) / \sum_{w=1 \text{ a } W} \exp(v^{WT} \cdot v_i)$$

dove v_i e v_o rappresentano rispettivamente i vettori di input e output associati alle parole, e W denota la dimensione del vocabolario. Questa formulazione induce una struttura geometrica dove la similarità coseno tra vettori di parole riflette la loro co-occorrenza contestuale.

La geometria emergente da questo processo di ottimizzazione esibisce proprietà algebriche notevoli, manifestate attraverso relazioni di parallelismo tra vettori. La celebre equazione vettoriale:

$$v(re) - v(uomo) + v(donna) \approx v(regina)$$

illustra come le relazioni semantiche possano essere codificate attraverso operazioni vettoriali lineari. Questa proprietà suggerisce che lo spazio di embedding non costituisce semplicemente un insieme disordinato di punti, ma possiede una struttura affine che rispetta le analogie concettuali.

Estensione agli Spazi Multilingue: Il Problema dell'Allineamento

La transizione dagli embedding monolingue a quelli multilingue introduce complessità geometriche sostanziali. Quando si addestrano modelli di embedding separatamente per diverse lingue, gli spazi vettoriali risultanti sono inizialmente incommensurabili: non esiste una corrispondenza naturale tra le basi vettoriali, le scale metriche o le orientazioni degli assi. Il problema centrale della geometria interlinguistica consiste quindi nell'identificare trasformazioni che allineino questi spazi preservando le strutture semantiche interne.

L'approccio pionieristico proposto da Mikolov e colleghi nel 2013 formula questo problema come ricerca di una trasformazione lineare ottimale. Dato un dizionario bilingue contenente n coppie di parole tradotte (x_i, y_i) dove x_i appartiene allo spazio di embedding della lingua sorgente X e y_i allo spazio della lingua target Y , si cerca una matrice di trasformazione W che minimizzi la distanza euclidea tra i vettori trasformati:

$$\min_W \sum_{i=1 \text{ a } n} \|W \cdot x_i - y_i\|^2$$

Questa formulazione ammette una soluzione analitica attraverso la decomposizione ai valori singolari (SVD). Ponendo $X = [x_1, x_2, \dots, x_n]$ e $Y = [y_1, y_2, \dots, y_n]$ come matrici contenenti i vettori di embedding, la matrice ottimale W può essere espressa come:

$$W = U \cdot V^T$$

dove U e V sono le matrici unitarie ottenute dalla SVD della matrice di cross-covarianza $Y \cdot X^T = U \cdot \Sigma \cdot V^T$.

Questa soluzione presenta una caratteristica geometrica fondamentale: la trasformazione ottimale è una rotazione ortogonale (o riflessione), preservando quindi le distanze e gli angoli nello spazio sorgente. La conservazione della struttura metrica garantisce che le relazioni semantiche codificate negli embedding monolingue vengano mantenute dopo la trasformazione.

Vincoli Ortogonali e Preservazione della Struttura

La scelta di limitare W ad essere una matrice ortogonale, ovvero tale che $W^T W = I$, dove I denota la matrice identità, deriva da considerazioni sia teoriche che empiriche. Dal punto di vista geometrico, le trasformazioni ortogonali appartengono al gruppo speciale ortogonale $SO(d)$ in dimensione d , costituendo isometrie che preservano prodotti scalari, norme e volumi. Questa proprietà è cruciale poiché le relazioni di similarità semantica negli embedding sono tipicamente quantificate attraverso la similarità coseno:

$$\cos(\theta) = (u \cdot v) / (||u|| \cdot ||v||)$$

Una trasformazione ortogonale W preserva esattamente questa quantità:

$$\cos(\theta') = (W \cdot u \cdot W \cdot v) / (||W \cdot u|| \cdot ||W \cdot v||) = (u \cdot v) / (||u|| \cdot ||v||) = \cos(\theta)$$

Questo assicura che se due parole nella lingua sorgente sono semanticamente correlate, la loro correlazione verrà mantenuta dopo la trasformazione nello spazio target.

Tuttavia, la restrizione alla pura ortogonalità si è rivelata talvolta eccessivamente stringente. Lavori successivi hanno esplorato generalizzazioni che permettono trasformazioni più flessibili preservando comunque le proprietà geometriche essenziali. Artetxe e colleghi nel 2016 hanno proposto un framework che decompone la trasformazione in una sequenza di operazioni geometriche elementari: normalizzazione, centratura, whitening ortogonale e rotazione finale.

La normalizzazione consiste nel rescalare tutti i vettori alla norma unitaria, proiettandoli sulla sfera unitaria $S^{(d-1)}$ in d dimensioni. Questa operazione elimina variazioni spurie nella magnitudine dei vettori che potrebbero derivare da differenze nelle frequenze lessicali tra le lingue. La centratura sottrae il centroide dei vettori, traslandolo nell'origine dello spazio. Il whitening ortogonale trasforma i dati in modo che la matrice di covarianza risultante sia proporzionale all'identità, eliminando correlazioni spurie tra dimensioni.

Geometria degli Spazi Cross-Linguistici: Isomorfismo Strutturale

Una scoperta empirica sorprendente emersa dalla ricerca sugli embedding multilingue riguarda il cosiddetto isomorfismo strutturale tra spazi linguistici. Nonostante le profonde differenze tipologiche, fonologiche e sintattiche tra lingue geneticamente non correlate, i loro spazi di embedding esibiscono strutture geometriche notevolmente simili quando addestrati su corpora comparabili. Questo fenomeno suggerisce l'esistenza di universali semantici che trascendono le particolarità superficiali delle lingue.

Søgaard e colleghi hanno quantificato questo isomorfismo attraverso metriche di similarità tra grafi di vicinanza. Costruendo per ciascuna lingua un grafo k -NN (k -nearest neighbors) dove i nodi rappresentano parole e gli archi connettono parole semanticamente vicine, hanno misurato la sovrapposizione strutturale tra grafi corrispondenti a lingue diverse dopo allineamento. I risultati indicano che per coppie di lingue europee, la sovrapposizione può raggiungere il 60-70%, un valore notevole considerando che le lingue sono state embeddate indipendentemente.

Questa proprietà può essere formalizzata attraverso il concetto di quasi-isometria tra spazi metrici. Due spazi metrici (X, d_x) e (Y, d_y) si dicono quasi-isometrici se esistono una funzione $f: X \rightarrow Y$ e costanti $\lambda \geq 1$, $\varepsilon \geq 0$ tali che:

$$(1/\lambda) \cdot d_x(x_1, x_2) - \varepsilon \leq d_y(f(x_1), f(x_2)) \leq \lambda \cdot d_x(x_1, x_2) + \varepsilon$$

per ogni coppia di punti x_1, x_2 in X . L'esistenza di trasformazioni lineari che approssimano quasi-isometrie tra spazi di embedding multilingue fornisce una giustificazione teorica per l'efficacia degli approcci basati su allineamento lineare.

Modelli Contestualizzati e Geometria Dinamica

L'avvento dei modelli di linguaggio contestualizzati, esemplificati da architetture come BERT (Bidirectional Encoder Representations from Transformers), ELMo (Embeddings from Language Models) e GPT (Generative Pre-trained Transformer), ha introdotto una complessità geometrica aggiuntiva. A differenza degli embedding statici di Word2Vec o GloVe, dove ciascuna parola è rappresentata da un singolo vettore indipendente dal contesto, i modelli contestualizzati generano rappresentazioni dinamiche che variano in funzione del contesto linguistico circostante.

Dal punto di vista geometrico, questo significa che anziché associare a ciascuna parola un punto fisso nello spazio di embedding, si ha una distribuzione di punti corrispondenti alle diverse realizzazioni contestuali. La polisemia, fenomeno per cui una parola possiede molteplici significati distinti, viene naturalmente catturata attraverso la dispersione geometrica delle rappresentazioni contestualizzate attorno a centroidi semantici differenti.

Per una parola polisemica w con k sensi distinti $s_1, s_2, \dots, s_k$, le rappresentazioni contestualizzate formano k cluster nello spazio di embedding. La struttura geometrica di questi cluster può essere analizzata attraverso tecniche di clustering gerarchico o metodi basati su mixture models. La distanza intra-cluster riflette la coerenza semantica di ciascun senso, mentre le distanze inter-cluster quantificano la dissimilarità tra i diversi significati.

Nel contesto multilingue, i modelli come mBERT (multilingual BERT) e XLM-R (Cross-lingual Language Model - RoBERTa) vengono addestrati congiuntamente su corpora in centinaia di lingue, inducendo uno spazio di rappresentazione intrinsecamente multilingue. La geometria di questi spazi presenta caratteristiche distintive: anziché richiedere un passo separato di allineamento post-hoc, le rappresentazioni cross-linguistiche emergono naturalmente dal processo di pre-addestramento attraverso il meccanismo dell'attenzione multi-testa.

Topologia degli Spazi di Embedding: Varietà e Curvatura

Un'analisi più profonda della geometria degli spazi di embedding rivela che questi non costituiscono semplici spazi euclidei piatti, ma esibiscono proprietà topologiche e geometriche più complesse assimilabili a varietà riemanniane immerse in spazi ambiente ad alta dimensionalità. Una varietà riemanniana è uno spazio che localmente assomiglia allo spazio euclideo ma può possedere una struttura globale non triviale e una curvatura intrinseca.

La curvatura gaussiana di una varietà bidimensionale in un punto p quantifica quanto la geometria locale si discosta dalla geometria euclidea piatta. Generalizzando a dimensioni superiori, si introduce il tensore di curvatura di Riemann, un oggetto matematico che codifica completamente le proprietà geometriche intrinseche della varietà. Per gli spazi di embedding linguistici, evidenze empiriche suggeriscono che certe regioni dello spazio presentano curvatura positiva (geometria sferica) mentre altre esibiscono curvatura negativa (geometria iperbolica).

In particolare, strutture gerarchiche come tassonomie e ontologie sembrano essere meglio rappresentate in spazi iperbolici piuttosto che euclidei. Lo spazio iperbolico di dimensione d , denotato $\mathbb{H}^d$, possiede curvatura costante negativa e la sua geometria permette di rappresentare alberi con distorsione logaritmica anziché polinomiale. Il modello di Poincaré, una delle rappresentazioni dello spazio iperbolico, utilizza la palla unitaria d -dimensionale con metrica:

$$ds^2 = 4 \cdot (dx_1^2 + dx_2^2 + \dots + dx_a^2) / (1 - x_1^2 - x_2^2 - \dots - x_a^2)^2$$

Questa metrica implica che la distanza tra punti cresce esponenzialmente man mano che ci si avvicina al bordo della palla, permettendo di embeddare strutture gerarchiche profonde in spazi di dimensionalità relativamente bassa.

Nel contesto multilingue, la scelta della geometria dello spazio di embedding può avere implicazioni significative per la qualità dell'allineamento cross-linguistico. Lingue con strutture sintattiche marcatamente gerarchiche, come quelle che esibiscono fenomeni di incorporazione nominale o ricorsione profonda, potrebbero beneficiare di rappresentazioni iperboliche che catturano più naturalmente le dipendenze annidate.

Manifold Learning e Riduzione Dimensionale

Gli spazi di embedding ad alta dimensionalità, sebbene espressivi, presentano sfide computazionali e statistiche. Il cosiddetto "curse of dimensionality" (maledizione della dimensionalità) implica che in spazi ad alta dimensione, i concetti di vicinanza e distanza diventano meno discriminativi: la maggior parte dei punti si trova a distanza approssimativamente uguale da qualsiasi punto di riferimento. Inoltre, la visualizzazione e l'interpretabilità degli embedding richiedono proiezioni in spazi di dimensionalità ridotta, tipicamente 2 o 3 dimensioni.

Il manifold learning costituisce una famiglia di tecniche non lineari per la riduzione dimensionale che assumono che i dati osservati ad alta dimensione giacciono su o vicino a una varietà di dimensionalità intrinseca molto inferiore. A differenza della Principal Component Analysis (PCA), che identifica direzioni di massima varianza attraverso trasformazioni lineari, i metodi di manifold learning come Isomap, Locally Linear Embedding (LLE) e t-SNE (t-distributed Stochastic Neighbor Embedding) preservano strutture geometriche non lineari.

L'algoritmo Isomap estende il classico Multi-Dimensional Scaling (MDS) sostituendo le distanze euclidee con distanze geodetiche sulla varietà. La distanza geodetica tra due punti è definita come la lunghezza del cammino più breve che li connette rimanendo sulla varietà. Operativamente, Isomap costruisce un grafo di vicinanza dove ciascun punto è connesso ai

suoi k vicini più prossimi, quindi calcola i cammini minimi su questo grafo approssimando le geodetiche, e infine applica MDS classico alla matrice delle distanze geodetiche.

Per gli spazi di embedding multilingue, Isomap può rivelare la struttura intrinseca degli spazi allineati, identificando dimensioni latenti che corrispondono a categorie semantiche universali. Ad esempio, una delle dimensioni principali potrebbe catturare la distinzione tra nomi concreti e astratti, mentre un'altra potrebbe codificare la valenza emotiva (positiva vs. negativa).

L'algoritmo t-SNE, ampiamente utilizzato per la visualizzazione di embedding, opera minimizzando la divergenza di Kullback-Leibler tra due distribuzioni di probabilità: una definita sulle coppie di punti nello spazio originale ad alta dimensione e una sulle coppie corrispondenti nello spazio ridotto. Le probabilità nello spazio originale sono proporzionali a gaussiane centrate sui punti:

$$p_{ij} = \exp(-\|x_i - x_j\|^2 / (2\sigma_i^2)) / \sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / (2\sigma_i^2))$$

mentre nello spazio ridotto si utilizza una distribuzione t di Student con un grado di libertà:

$$q_{ij} = (1 + \|y_i - y_j\|^2)^{-1} / \sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}$$

La scelta della distribuzione t è cruciale: le sue code pesanti permettono di rappresentare distanze moderate nello spazio originale come distanze più grandi nello spazio ridotto, evitando il "crowding problem" dove punti distinti vengono collassati insieme.

Geometria delle Rappresentazioni Sintattiche

Oltre alle rappresentazioni lessicali, la geometria degli spazi di embedding si estende alle strutture sintattiche. I modelli neurali per il parsing sintattico, come i parser basati su transizioni o su grafici, operano costruendo rappresentazioni vettoriali di configurazioni sintattiche intermedie. Queste rappresentazioni formano traiettorie geometriche nello spazio di embedding mentre il parser processa incrementalmente la frase.

Le trasformazioni sintattiche cross-linguistiche, come la permutazione dell'ordine delle parole tra lingue SVO (Soggetto-Verbo-Oggetto) e SOV (Soggetto-Oggetto-Verbo), possono essere modellate come trasformazioni geometriche sulle traiettorie di parsing. Un approccio promettente consiste nell'apprendere mappe differenziabili che trasformano la traiettoria di parsing in una lingua nella corrispondente traiettoria in un'altra lingua, preservando la struttura di dipendenza sottostante.

Formalmente, data una sequenza di configurazioni di parsing $C_1, C_2, \dots, C_n$ nella lingua sorgente, ciascuna rappresentata da un vettore c_i nello spazio di embedding, si cerca una trasformazione $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^d$ tale che le configurazioni trasformate $\phi(c_1), \phi(c_2), \dots, \phi(c_n)$ corrispondano a una sequenza valida di configurazioni nella lingua target. La funzione ϕ può essere parametrizzata attraverso reti neurali profonde, e appresa mediante supervisione o attraverso tecniche di reinforcement learning.

Conclusioni del Fondamento Teorico

Questa prima parte ha stabilito i fondamenti geometrici necessari per comprendere la traduzione automatica come problema di trasformazione tra spazi vettoriali strutturati. Abbiamo esplorato la genesi degli spazi di embedding monolingue, il problema dell'allineamento cross-linguistico attraverso trasformazioni ortogonali, l'isomorfismo strutturale tra lingue diverse, la transizione verso rappresentazioni contestualizzate, le proprietà topologiche degli spazi come varietà riemanniane, le tecniche di manifold learning per la riduzione dimensionale, e infine la geometria delle rappresentazioni sintattiche.

Queste fondamenta teoriche preparano il terreno per un'analisi approfondita delle trasformazioni geometriche specifiche utilizzate nell'allineamento cross-linguistico, delle proiezioni tra varietà linguistiche, della geometria delle divergenze tipologiche, e dell'applicazione del geometric deep learning alla traduzione automatica neurale, temi che verranno sviluppati nelle parti successive di questo trattato.

Trasformazioni Geometriche e Allineamenti Cross-Linguistici

Rotazioni Ortogonali e il Problema di Procrustes

L'allineamento di spazi di embedding cross-linguistici può essere formulato elegantemente come un problema di Procrustes ortogonale, un classico problema di ottimizzazione in geometria computazionale che prende il nome dal mitologico brigante greco Procruste, il quale adattava le vittime al suo letto attraverso stiramenti o amputazioni. Nel nostro contesto, cerchiamo la trasformazione ortogonale ottimale che "adatta" uno spazio di embedding a un altro minimizzando la distorsione complessiva.

Data una matrice X di dimensioni $n \times d$ contenente n vettori di embedding della lingua sorgente in d dimensioni, e una matrice Y della stessa forma contenente i corrispondenti vettori nella lingua target, il problema di Procrustes ortogonale consiste nel trovare la matrice di rotazione R appartenente al gruppo ortogonale speciale $SO(d)$ che minimizza la norma di Frobenius della differenza:

$$\min_{R \in SO(d)} \|X \cdot R - Y\|_F^2 = \min_{R \in SO(d)} \text{tr}((X \cdot R - Y)^T \cdot (X \cdot R - Y))$$

dove $\text{tr}(\cdot)$ denota la traccia di una matrice e $\|\cdot\|_F$ la norma di Frobenius. Espandendo il termine quadratico e semplificando, questa espressione diventa:

$$\min_{R \in SO(d)} \text{tr}(X^T X \cdot R^T R) - 2 \cdot \text{tr}(X^T Y \cdot R) + \text{tr}(Y^T Y)$$

Poiché R è ortogonale, si ha $R^T R = I$, e i termini $\text{tr}(X^T X)$ e $\text{tr}(Y^T Y)$ sono costanti indipendenti da R . Il problema si riduce quindi alla massimizzazione di:

$$\max_{R \in SO(d)} \text{tr}(X^T Y \cdot R) = \max_{R \in SO(d)} \text{tr}(R \cdot X^T Y)$$

utilizzando la proprietà ciclica della traccia $\text{tr}(AB) = \text{tr}(BA)$.

La soluzione ottimale può essere ottenuta attraverso la decomposizione ai valori singolari della matrice di cross-correlazione $M = Y^T X$. Scrivendo $M = U \Sigma V^T$ dove U e V sono matrici

ortogonali e Σ è diagonale con elementi non negativi (i valori singolari), la rotazione ottimale è data da:

$$R = V \cdot U^T$$

Questa soluzione garantisce che $\det(R) = \pm 1$. Se $\det(R) = -1$, si ha una riflessione anziché una pura rotazione. Per ottenere una rotazione propria (con $\det(R) = +1$), è necessario modificare la soluzione correggendo il segno dell'ultimo vettore singolare quando il determinante risulta negativo.

La bellezza geometrica di questa soluzione risiede nella sua interpretazione: la decomposizione SVD identifica un sistema di coordinate principali per entrambi gli spazi, e la trasformazione ottimale consiste semplicemente nell'allineare questi sistemi. I valori singolari $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d$ quantificano quanto bene gli spazi possono essere allineati lungo ciascuna direzione principale: valori singolari prossimi a 1 indicano un allineamento quasi perfetto, mentre valori piccoli segnalano direzioni lungo le quali gli spazi divergono significativamente.

Trasformazioni Affini e Rescaling Anisotropo

Sebbene le rotazioni ortogonali preservino esattamente le distanze e gli angoli, imponendo una rigidità che può risultare eccessivamente restrittiva in alcuni contesti, è possibile rilassare questo vincolo ammettendo trasformazioni affini più generali. Una trasformazione affine ha la forma:

$$y = A \cdot x + b$$

dove A è una matrice $d \times d$ arbitraria e b è un vettore di traslazione. Questa famiglia include rotazioni, riflessioni, rescaling isotropi e anisotropi, shearing e traslazioni.

Nel contesto dell'allineamento cross-linguistico, la componente di traslazione b corrisponde a uno shift globale tra i centroidi degli spazi di embedding. Se gli spazi sono stati preventivamente centrati sottraendo le medie, questo termine diventa nullo. Il termine più interessante è la matrice A , che può essere decomposta mediante la decomposizione polare $A = R \cdot S$ dove R è ortogonale e S è simmetrica definita positiva.

La matrice simmetrica S può essere ulteriormente diagonalizzata come $S = Q^T \Lambda Q$ dove Q è ortogonale e Λ è diagonale con elementi positivi $\lambda_1, \lambda_2, \dots, \lambda_d$ rappresentanti i fattori di rescaling lungo gli assi principali. Questa decomposizione rivela che una trasformazione affine generale può essere vista come sequenza di: rotazione Q , rescaling anisotropo Λ lungo gli assi coordinati, e rotazione finale R .

Il rescaling anisotropo è particolarmente rilevante per modellare differenze tipologiche tra lingue. Ad esempio, lingue con sistemi di genere grammaticale ricco potrebbero richiedere più dimensioni per codificare distinzioni di genere rispetto a lingue senza genere grammaticale. Un fattore di rescaling $\lambda_i > 1$ lungo una particolare direzione i espanderebbe quella dimensione semantica, mentre $\lambda_i < 1$ la contrarrebbe.

L'apprendimento di trasformazioni affini generali può essere formulato come problema di regressione lineare multivariata. Data la collezione di coppie di embedding allineati $\{(x_i, y_i)\}_{i=1}^n$, si minimizza:

$$\min_{A,b} \sum_{i=1}^n \|A \cdot x_i + b - y_i\|^2$$

Questo problema ammette una soluzione analitica in forma chiusa che può essere derivata annullando i gradienti rispetto ad A e b . Tuttavia, senza regolarizzazione, la matrice A appresa può risultare mal condizionata o sovraadattata ai dati di training.

Per mitigare questi problemi, si introducono tipicamente termini di regolarizzazione che penalizzano la complessità della trasformazione. Una scelta comune è la regolarizzazione di Tikhonov (o ridge regression):

$$\min_{A,b} \sum_{i=1}^n \|A \cdot x_i + b - y_i\|^2 + \alpha \cdot \|A\|_F^2$$

dove $\alpha > 0$ è un iperparametro che controlla l'intensità della regolarizzazione. Valori elevati di α forzano A verso la matrice identità, riducendo a una semplice traslazione, mentre $\alpha \rightarrow 0$ recupera la soluzione non regolarizzata.

Approcci Non Supervisionati e Adversarial Training

Un limite significativo dei metodi descritti finora è la loro dipendenza da dizionari bilingui o parallelismi lessicali per stabilire le corrispondenze iniziali tra gli spazi. Nella pratica, tali risorse possono essere costose da ottenere, specialmente per coppie di lingue a basse risorse o per domini specializzati. Questo ha motivato lo sviluppo di approcci non supervisionati che allineano spazi di embedding senza alcun segnale di supervisione esplicito.

L'idea fondamentale di questi metodi si basa sull'osservazione che, nonostante l'assenza di allineamento, gli spazi di embedding monolingue esibiscono strutture statistiche simili che possono essere sfruttate come ancora per l'allineamento. In particolare, le distribuzioni dei vettori di embedding tendono a essere approssimativamente invarianti sotto trasformazioni ortogonali, suggerendo un criterio di allineamento basato sul matching distributivo.

L'approccio MUSE (Multilingual Unsupervised and Supervised Embeddings), proposto da Conneau e colleghi, impiega un framework adversarial ispirato alle Generative Adversarial Networks (GANs). L'architettura consiste di due componenti principali: un generatore che applica una trasformazione agli embedding della lingua sorgente, e un discriminatore che cerca di distinguere tra embedding trasformati della lingua sorgente ed embedding nativi della lingua target.

Formalmente, sia W la matrice di trasformazione parametrizzante il generatore. Il discriminatore è una rete neurale $D(\cdot)$ che produce un punteggio scalare indicante la probabilità che un embedding provenga dalla lingua target. L'obiettivo del generatore è ingannare il discriminatore rendendo indistinguibili gli embedding trasformati $W \cdot x$ da quelli target y , mentre il discriminatore cerca di massimizzare la sua accuratezza discriminativa.

Questo gioco minimax può essere espresso attraverso la funzione obiettivo:

$$\min W \max D E_x[\log D(W \cdot x)] + E_y[\log(1 - D(y))]$$

dove E_x denota il valore atteso rispetto alla distribuzione degli embedding sorgente e E_y rispetto alla distribuzione target. L'addestramento procede alternando aggiornamenti del discriminatore (massimizzazione rispetto a D mantenendo W fisso) e del generatore (minimizzazione rispetto a W mantenendo D fisso).

Un aspetto cruciale è che la trasformazione W viene vincolata ad essere ortogonale durante l'ottimizzazione. Questo vincolo può essere imposto utilizzando la parametrizzazione della varietà di Stiefel mediante ottimizzazione su varietà riemanniane, oppure attraverso tecniche di proiezione che, dopo ogni aggiornamento gradiente, proiettano W sulla varietà ortogonale più vicina.

Per raffinare ulteriormente l'allineamento ottenuto attraverso l'addestramento adversarial, MUSE impiega un passo di raffinamento iterativo. Si costruisce un dizionario sintetico selezionando per ciascun embedding sorgente il suo vicino più prossimo nello spazio target dopo la trasformazione. Questo dizionario viene quindi utilizzato per risolvere il problema di Procrustes ortogonale, ottenendo una nuova trasformazione W' . Il processo si itera fino a convergenza.

Optimal Transport e Wasserstein Alignment

Un approccio alternativo e matematicamente rigoroso all'allineamento non supervisionato si basa sulla teoria del trasporto ottimale. Il trasporto ottimale, originariamente formulato da Gaspard Monge nel 1781 e successivamente rilassato da Leonid Kantorovich, affronta il problema di come trasportare massa da una distribuzione sorgente a una distribuzione target minimizzando il costo complessivo di trasporto.

Nel contesto degli embedding cross-linguistici, si interpretano le distribuzioni empiriche degli embedding nelle due lingue come misure di probabilità μ e ν su $\mathbb{R}^d$. Il problema di Monge-Kantorovich consiste nel trovare un piano di trasporto π , ovvero una misura di probabilità congiunta su $\mathbb{R}^d \times \mathbb{R}^d$ con marginali μ e ν , che minimizzi il costo di trasporto totale:

$$\min \pi \in \Pi(\mu, \nu) \iint c(x, y) d\pi(x, y)$$

dove $c(x, y)$ è una funzione di costo, tipicamente la distanza euclidea al quadrato $c(x, y) = \|x - y\|^2$, e $\Pi(\mu, \nu)$ denota l'insieme di tutti i piani di trasporto con marginali prescritte.

La distanza di Wasserstein tra μ e ν è definita come la radice quadrata del costo di trasporto ottimale:

$$W_2(\mu, \nu) = \sqrt{(\min \pi \in \Pi(\mu, \nu) \iint \|x - y\|^2 d\pi(x, y))}$$

Questa distanza induce una metrica nello spazio delle misure di probabilità e possiede proprietà geometriche favorevoli, come la convessità lungo curve geodetiche nello spazio delle misure.

Per il caso di distribuzioni empiriche discrete $\mu = (1/n) \sum_i \delta_{x_i}$ e $\nu = (1/m) \sum_j \delta_{y_j}$, il problema di trasporto ottimale si riduce a un problema di programmazione lineare.

Definendo la matrice di costo C dove $C_{ij} = \|x_i - y_j\|^2$, e una matrice di piano di trasporto P dove P_{ij} rappresenta la massa trasportata da x_i a y_j , il problema diventa:

$$\min P \sum_i \sum_j C_{ij} \cdot P_{ij}$$

soggetto a: $\sum_j P_{ij} = 1/n$ per ogni i , $\sum_i P_{ij} = 1/m$ per ogni j , $P_{ij} \geq 0$

Questo può essere risolto efficientemente usando l'algoritmo del simplesso o metodi di punto interno. Tuttavia, per dataset di grandi dimensioni con migliaia o milioni di parole, la complessità computazionale diventa proibitiva.

L'algoritmo di Sinkhorn fornisce un'alternativa computazionalmente efficiente introducendo un termine di regolarizzazione entropica. Il problema regolarizzato diventa:

$$\min P \sum_i \sum_j C_{ij} \cdot P_{ij} - (1/\lambda) \cdot H(P)$$

dove $H(P) = -\sum_i \sum_j P_{ij} \cdot \log(P_{ij})$ è l'entropia della matrice di trasporto e $\lambda > 0$ controlla l'intensità della regolarizzazione. La soluzione può essere ottenuta mediante iterazioni di Sinkhorn che alternano normalizzazioni delle righe e colonne della matrice $K = \exp(-\lambda \cdot C)$:

$$u^{k+1}_i = (1/n) / \sum_j K_{ij} \cdot v^k_j \quad v^{k+1}_j = (1/m) / \sum_i K_{ij} \cdot u^{k+1}_i$$

con la soluzione finale $P_{ij} = u_i \cdot K_{ij} \cdot v_j$. Questo algoritmo converge geometricamente ed è facilmente parallelizzabile su GPU.

Nel contesto dell'allineamento cross-linguistico, Grave e colleghi hanno proposto di utilizzare il trasporto ottimale per costruire una trasformazione tra spazi che rispetti la struttura distributiva. Aniché cercare una singola trasformazione globale, si identifica per ciascun embedding sorgente il suo target ottimale secondo il piano di trasporto, quindi si apprende una trasformazione (tipicamente una rete neurale) che interpola questi mapping.

Geometria delle Proiezioni Cross-Linguistiche

Oltre alle trasformazioni lineari e affini, un approccio concettualmente diverso consiste nel costruire uno spazio di embedding multilingue condiviso nel quale tutte le lingue vengono proiettate. Questa prospettiva sposta il focus dall'allineamento post-hoc di spazi monolingue alla costruzione diretta di uno spazio universale.

I metodi di proiezione possono essere categorizzati in approcci basati su dati paralleli e approcci basati su comparabilità. I primi richiedono frasi allineate tra le lingue, mentre i secondi operano su corpora monolingue comparabili (documenti su argomenti simili in lingue diverse) senza allineamento esplicito a livello di frase.

Un approccio influente basato su dati paralleli è quello dei Bilingual Word Embeddings tramite autoencoders. Si addestrano due encoder E_s e E_t che mappano rispettivamente gli embedding monolingue sorgente e target in uno spazio condiviso di dimensionalità potenzialmente ridotta, e due decoder D_s e D_t che ricostruiscono gli embedding originali:

$$x_s \rightarrow E_s(x_s) \rightarrow D_s(E_s(x_s)) \approx x_s \quad x_t \rightarrow E_t(x_t) \rightarrow D_t(E_t(x_t)) \approx x_t$$

Per coppie di parole allineate (x_i, x_j) provenienti da un dizionario bilingue o da allineamenti automatici, si impone un vincolo addizionale che le rappresentazioni nello spazio condiviso siano identiche o molto vicine:

$$\|E_i(x_i) - E_j(x_j)\| \approx 0$$

La funzione di perdita combinata diventa:

$$L = \sum_i \|x_i - D_i(E_i(x_i))\|^2 + \sum_j \|x_j - D_j(E_j(x_j))\|^2 + \gamma \cdot \sum_{i,j} \|E_i(x_i) - E_j(x_j)\|^2$$

dove il primo termine è la perdita di ricostruzione per la lingua sorgente, il secondo per la lingua target, e il terzo è il termine di allineamento cross-linguistico pesato da $\gamma > 0$.

Questo framework può essere esteso a più di due lingue definendo un encoder per ciascuna lingua E_l per $l \in \{l_1, l_2, \dots, l_n\}$, tutti proiettanti in un unico spazio condiviso. Gli autoencoders possono essere implementati come semplici trasformazioni lineari (nel qual caso si riducono essenzialmente a PCA multilingue) oppure come reti neurali profonde non lineari che possono catturare strutture più complesse.

Canonical Correlation Analysis Multilingue

La Canonical Correlation Analysis (CCA) è una tecnica statistica classica per identificare correlazioni lineari tra due insiemi di variabili. Dati due insiemi di vettori X e Y , la CCA cerca direzioni canoniche w_x e w_y che massimizzano la correlazione tra le proiezioni:

$$\max_{w_x, w_y} \text{corr}(Xw_x, Yw_y) = \max_{w_x, w_y} (w_x^T X^T Y w_y) / \sqrt{(w_x^T X^T X w_x)(w_y^T Y^T Y w_y)}$$

Questo problema può essere risolto mediante autodecomposizione generalizzata. Le prime k coppie di direzioni canoniche definiscono k componenti di massima correlazione, fornendo una rappresentazione condivisa di dimensionalità k .

Nel contesto multilingue, la CCA può essere applicata agli embedding di coppie di lingue per identificare sottospazi correlati. Faruqui e Dyer hanno proposto una estensione multilingue della CCA che generalizza a più di due lingue simultaneamente. L'obiettivo è trovare proiezioni W_l per ciascuna lingua l che massimizzino la correlazione media tra tutte le coppie di lingue:

$$\max \{W_l\} (1/(L(L-1)/2)) \sum_{l < l'} \text{corr}(X_l W_l, X_{l'} W_{l'})$$

dove L è il numero totale di lingue e X_l denota la matrice di embedding per la lingua l .

Una proprietà attraente della CCA è che le proiezioni ottimali sono invarianti rispetto a trasformazioni invertibili applicate separatamente a ciascuno spazio di input. Questo significa che la CCA identifica correlazioni intrinseche indipendentemente da rescaling o rotazioni arbitrarie degli embedding originali.

Dal punto di vista geometrico, la CCA può essere interpretata come ricerca di sottospazi che sono massimamente allineati tra le lingue. I vettori canonici w_l definiscono basi ortonormali per questi sottospazi, e le correlazioni canoniche quantificano quanto bene i sottospazi si allineano lungo ciascuna direzione.

Trasformazioni Composte e Cascade di Allineamento

In scenari che coinvolgono molte lingue, può essere computazionalmente inefficiente o addirittura impossibile apprendere allineamenti diretti tra tutte le coppie. Un'alternativa consiste nell'organizzare le lingue in una struttura gerarchica o grafo, e comporre trasformazioni lungo cammini nel grafo.

Supponiamo di avere appreso trasformazioni $W_{s \rightarrow t}$ che mappano embedding dalla lingua s alla lingua t , e $W_{t \rightarrow u}$ dalla lingua t alla lingua u . La trasformazione composta dalla lingua s alla lingua u può essere ottenuta come:

$$W_{s \rightarrow u} = W_{t \rightarrow u} \cdot W_{s \rightarrow t}$$

Se tutte le trasformazioni sono ortogonali, anche la composizione rimane ortogonale poiché il prodotto di matrici ortogonali è ortogonale. Tuttavia, la qualità dell'allineamento può degradare con il numero di composizioni a causa dell'accumulo di errori.

Per mitigare questo problema, si può selezionare una lingua pivot centrale (tipicamente l'inglese, data la sua abbondanza di risorse) e apprendere trasformazioni dirette da tutte le altre lingue verso il pivot. Gli allineamenti indiretti tra coppie di lingue non-pivot vengono quindi ottenuti transitivamente attraverso il pivot.

Un approccio più sofisticato consiste nell'apprendere simultaneamente tutte le trasformazioni in modo che soddisfino vincoli di consistenza. Dati tre lingue s , t , u con trasformazioni $W_{s \rightarrow t}$, $W_{s \rightarrow u}$, $W_{t \rightarrow u}$, si può imporre che la trasformazione diretta sia approssimativamente uguale alla composizione:

$$W_{s \rightarrow u} \approx W_{t \rightarrow u} \cdot W_{s \rightarrow t}$$

Questo vincolo può essere incorporato nella funzione obiettivo durante l'apprendimento mediante un termine di regolarizzazione:

$$L_{\text{consist}} = \sum_{s,t,u} \|W_{s \rightarrow u} - W_{t \rightarrow u} \cdot W_{s \rightarrow t}\|_F^2$$

La minimizzazione congiunta dell'errore di allineamento per ciascuna coppia e del termine di consistenza produce un insieme coerente di trasformazioni che rispettano la struttura di gruppo.

Allineamento Dinamico e Contestuale

Gli approcci discussi finora operano su embedding statici, dove ciascuna parola è rappresentata da un singolo vettore indipendente dal contesto. Come menzionato precedentemente, i modelli contestualizzati generano rappresentazioni dinamiche che variano con il contesto. L'allineamento di spazi contestualizzati introduce sfide aggiuntive poiché non esiste più una corrispondenza uno-a-uno tra parole e vettori.

Un approccio naturale consiste nell'aggregare le rappresentazioni contestualizzate di una parola attraverso i suoi diversi contesti, ottenendo un embedding medio o prototipico, quindi

applicare le tecniche di allineamento standard. Tuttavia, questa aggregazione può perdere informazioni cruciali sulla struttura dei cluster semantici.

Metodi più sofisticati apprendono allineamenti che variano con il contesto. Schuster e colleghi hanno proposto di apprendere una trasformazione dipendente dal contesto $W(c)$ parametrizzata come funzione del contesto c . In pratica, $W(c)$ può essere implementato come una rete neurale che riceve in input la rappresentazione del contesto e produce una matrice di trasformazione.

Un'alternativa consiste nell'utilizzare tecniche di matching ottimale per allineare i cluster di sensi. Per una parola polisemica con k sensi nella lingua sorgente e h sensi nella lingua target, si costruiscono cluster di rappresentazioni contestualizzate per ciascun senso, quindi si risolve un problema di matching bipartito per associare i cluster sorgente ai cluster target basandosi su una misura di similarità inter-cluster.

Invarianza e Equivarianza nelle Trasformazioni

Un principio fondamentale nella progettazione di trasformazioni cross-linguistiche è il rispetto di simmetrie e invarianze. Una trasformazione si dice invariante rispetto a una trasformazione g se $T(g(x)) = T(x)$, ed equivariante se $T(g(x)) = g(T(x))$. Nella pratica degli embedding multilingue, desideriamo che certe proprietà geometriche siano preservate.

Ad esempio, se due parole sono sinonimi nella lingua sorgente, ci aspettiamo che le loro traduzioni siano sinonimi nella lingua target. Questo implica che la trasformazione dovrebbe preservare la similarità coseno alta. Formalmente, se $\text{sim}(x_1, x_2) \approx 1$ nella lingua sorgente, allora $\text{sim}(T(x_1), T(x_2)) \approx 1$ nella lingua target.

Le trasformazioni ortogonali soddisfano questa proprietà poiché preservano esattamente i prodotti scalari. Trasformazioni più generali devono essere progettate con attenzione per non violare le invarianze semantiche desiderate.

Un framework matematico per formalizzare queste considerazioni si basa sulla teoria delle rappresentazioni di gruppi. Se G è un gruppo di simmetrie che vogliamo preservare (ad esempio, rotazioni o permutazioni), possiamo richiedere che la trasformazione T commuti con l'azione di G :

$$T(g \cdot x) = g \cdot T(x) \text{ per ogni } g \in G$$

Questa condizione di equivarianza garantisce che la struttura di simmetria viene preservata attraverso la trasformazione.

Metriche di Valutazione per l'Allineamento

La qualità di un allineamento cross-linguistico può essere valutata attraverso diverse metriche, ciascuna catturando aspetti differenti della correttezza geometrica e semantica.

La metrica più diretta è l'accuratezza nel task di traduzione lessicale: dato un insieme di test di coppie di parole tradotte, si misura la percentuale di casi in cui la traduzione corretta

appare tra i k vicini più prossimi dell'embedding sorgente trasformato nello spazio target. Tipicamente si riportano le accuratze $P@1$ (precisione al primo vicino), $P@5$, e $P@10$.

Una metrica più raffinata è il Mean Reciprocal Rank (MRR), che considera la posizione della traduzione corretta nella lista ordinata di vicini:

$$\text{MRR} = (1/n) \sum_{i=1}^n 1/\text{rank}_i$$

dove rank_i è la posizione (a partire da 1) della traduzione corretta per la i -esima parola di test.

Per valutare la preservazione della struttura geometrica, si può misurare la correlazione di Spearman tra distanze nello spazio sorgente e distanze corrispondenti nello spazio target dopo la trasformazione. Una correlazione alta indica che la struttura metrica viene preservata.

Un'altra metrica rilevante è l'isomorfismo del grafo di vicinanza: si costruisce un grafo k -NN per un insieme di parole nello spazio sorgente, si applica la trasformazione ottenendo un grafo nello spazio target, quindi si misura la sovrapposizione degli archi (percentuale di archi conservati). Un isomorfismo perfetto corrisponde a una sovrapposizione del 100%.

Conclusioni sulle Trasformazioni Geometriche

Questa seconda parte ha esplorato in profondità le trasformazioni geometriche che costituiscono il cuore dell'allineamento cross-linguistico. Abbiamo analizzato le rotazioni ortogonali attraverso il problema di Procrustes, le trasformazioni affini con rescaling anisotropo, gli approcci non supervisionati basati su adversarial training, il trasporto ottimale e le distanze di Wasserstein, le proiezioni in spazi condivisi mediante autoencoders, la Canonical Correlation Analysis multilingue, le trasformazioni composte e le cascate di allineamento, l'allineamento dinamico per rappresentazioni contestualizzate, i principi di invarianza ed equivarianza, e le metriche di valutazione.

Questi strumenti matematici e computazionali forniscono un arsenale ricco per affrontare la sfida dell'allineamento semantico tra lingue, ciascuno con i propri trade-off in termini di espressività, complessità computazionale, requisiti di supervisione, e robustezza. La scelta dell'approccio appropriato dipende dalle caratteristiche specifiche delle lingue coinvolte, dalla disponibilità di risorse parallelismo lessicale, e dagli obiettivi applicativi.

Proiezioni tra Varietà Linguistiche e Geometria delle Divergenze Tipologiche

Varietà Linguistiche come Manifold Differenziabili

La concezione degli spazi di embedding linguistici come varietà riemanniane apre prospettive teoriche profonde sulla natura geometrica del linguaggio. Una varietà differenziabile M di dimensione d è uno spazio topologico che localmente assomiglia a $\mathbb{R}^d$, dotato di una struttura di classe C^∞ che permette di definire operazioni di calcolo

differenziale. Quando la varietà è equipaggiata con un tensore metrico g che definisce prodotti scalari negli spazi tangenti in modo liscio, si ottiene una varietà riemanniana.

Nel contesto degli embedding linguistici, la varietà M rappresenta lo spazio effettivo occupato dagli embedding, che tipicamente è una sottovarietà di dimensione intrinseca molto inferiore rispetto allo spazio ambiente $\mathbb{R}^D$ in cui gli embedding sono formalmente rappresentati. Questa osservazione è supportata da evidenze empiriche che mostrano come gli embedding, sebbene formalmente vivano in spazi di 300-1000 dimensioni, possiedano una dimensionalità intrinseca significativamente minore, spesso nell'ordine di 100-200 dimensioni.

Il tensore metrico g in un punto p della varietà assegna a ciascuna coppia di vettori tangenti $u, v \in T_p M$ (lo spazio tangente in p) un prodotto scalare $g(u, v)$. Questo permette di definire lunghezze di curve, angoli tra direzioni, volumi di regioni, e crucialmente, geodetiche. Una geodetica $\gamma(t)$ è una curva che localmente minimizza la distanza tra punti, generalizzando il concetto di linea retta allo spazio curvo.

Le equazioni delle geodetiche su una varietà riemanniana sono date dalle equazioni di Euler-Lagrange derivanti dalla minimizzazione del funzionale di lunghezza:

$$d^2\gamma^i/dt^2 + \sum_j \Gamma^i_{jk} (d\gamma^j/dt)(d\gamma^k/dt) = 0$$

dove Γ^i_{jk} sono i simboli di Christoffel, che dipendono dalle derivate del tensore metrico e codificano la curvatura della varietà. In uno spazio euclideo piatto, i simboli di Christoffel sono nulli e le geodetiche sono semplicemente linee rette. In spazi curvi, le geodetiche deviano da percorsi rettilinei in modi che riflettono la geometria intrinseca.

Per gli spazi di embedding linguistici, la metrica può essere definita in modi diversi a seconda dell'interpretazione desiderata. Una scelta naturale è ereditare la metrica euclidea dallo spazio ambiente:

$$g(u, v) = u^T \cdot v$$

Tuttavia, questa scelta può non catturare strutture geometriche più sottili. Approcci alternativi definiscono metriche indotte da informazioni di Fisher o da kernel definiti sulle distribuzioni di contesto delle parole.

Curvatura Intrinseca e Sezionale

La curvatura di una varietà riemanniana quantifica la deviazione locale della geometria da quella euclidea piatta. Per varietà bidimensionali (superfici), la curvatura gaussiana K in un punto p è un singolo scalare che descrive completamente la geometria locale. Per superfici immerse in $\mathbb{R}^3$, la curvatura gaussiana è data dal prodotto delle curvature principali κ_1 e κ_2 :

$$K = \kappa_1 \cdot \kappa_2$$

Una sfera di raggio r ha curvatura costante positiva $K = 1/r^2$, un piano ha $K = 0$, e una superficie a sella (iperbolica) ha $K < 0$.

Per varietà di dimensione superiore, la curvatura è descritta dal tensore di Riemann $R^i{}_{jkl}$, un oggetto con 4 indici che cattura come i vettori cambiano quando vengono trasportati parallelamente lungo curve chiuse. Dal tensore di Riemann si derivano diverse nozioni di curvatura scalare.

La curvatura sezionale $K(\sigma)$ associata a un piano bidimensionale $\sigma \subset T_x M$ nello spazio tangente generalizza la curvatura gaussiana. È definita come:

$$K(\sigma) = R(u, v, v, u) / (g(u, u)g(v, v) - g(u, v)^2)$$

dove u, v sono vettori linearmente indipendenti che generano il piano σ , e R denota il tensore di Riemann contratto appropriatamente.

Nel contesto linguistico, la curvatura sezionale può rivelare strutture semantiche profonde. Regioni di curvatura positiva corrispondono a cluster semantici compatti dove i concetti sono densamente interconnessi, simile alla geometria di una sfera. Regioni di curvatura negativa corrispondono a strutture gerarchiche espansive, come quelle osservabili in ontologie e tassonomie, dove lo spazio disponibile cresce esponenzialmente con la distanza dal punto di riferimento.

Nickel e Kiela hanno dimostrato empiricamente che embedding di gerarchie lessicali come WordNet esibiscono curvatura negativa sistematica quando analizzati attraverso tecniche di ricostruzione geometrica. Questo ha motivato l'esplorazione di embedding iperbolici che catturano naturalmente queste strutture gerarchiche.

Embedding Iperbolici per Strutture Gerarchiche

Lo spazio iperbolico di dimensione d , denotato $\mathbb{H}^d$, è una varietà riemanniana a curvatura costante negativa -1 . Esistono diversi modelli equivalenti di geometria iperbolica, tra cui il modello di Poincaré (disco o semispazio superiore), il modello di Klein-Beltrami, e il modello iperboloide (o di Lorentz).

Il modello del disco di Poincaré rappresenta $\mathbb{H}^d$ come la palla unitaria aperta $B^d = \{x \in \mathbb{R}^d : \|x\| < 1\}$ equipaggiata con la metrica:

$$ds^2 = 4 \cdot (dx_1^2 + \dots + dx_d^2) / (1 - \|x\|^2)^2$$

Questa metrica diverge esponenzialmente quando ci si avvicina al bordo $\|x\| \rightarrow 1$, conferendo allo spazio una "capacità infinita" vicino al boundary. Le geodetiche nel disco di Poincaré sono archi di cerchio ortogonali al bordo, o diametri quando passano per l'origine.

La distanza iperbolica tra due punti u, v nel disco di Poincaré è data da:

$$d_{\mathbb{H}}(u, v) = \operatorname{arcosh}(1 + 2 \cdot \|u - v\|^2 / ((1 - \|u\|^2)(1 - \|v\|^2)))$$

dove arcosh è il coseno iperbolico inverso. Questa distanza cresce logaritmicamente rispetto al numero di nodi in un albero perfettamente bilanciato, permettendo di embeddare alberi binari completi di profondità h con distorsione $O(1)$ in uno spazio iperbolico di dimensione 2, mentre uno spazio euclideo richiederebbe dimensionalità esponenziale in h .

Per apprendere embedding iperbolici, è necessario adattare gli algoritmi di ottimizzazione per operare sulla varietà riemanniana $\mathbb{H}^d$ anziché nello spazio euclideo. L'ottimizzazione riemanniana generalizza la discesa del gradiente sostituendo il gradiente euclideo con il gradiente riemanniano e gli aggiornamenti lineari con aggiornamenti lungo geodetiche.

Dato un punto corrente $x \in \mathbb{H}^d$ e un gradiente euclideo ∇f , il gradiente riemanniano è:

$$\text{grad } f = (1 - \|x\|^2)^2 / 4 \cdot \nabla f$$

che compensa la distorsione metrica. L'aggiornamento viene quindi eseguito mediante il mappa esponenziale che trasporta un vettore tangente $v \in T_x M$ lungo la geodetica per tempo t :

$$x_{\text{new}} = \exp_x(-\eta \cdot \text{grad } f)$$

dove η è il learning rate e $\exp_x$ è la mappa esponenziale in x . Nel disco di Poincaré, la mappa esponenziale ha una forma analitica esplicita che può essere calcolata efficientemente.

Allineamento tra Spazi Iperbolici

L'allineamento di embedding iperbolici cross-linguistici introduce complessità aggiuntive rispetto al caso euclideo. Le trasformazioni che preservano la struttura iperbolica appartengono al gruppo di Lorentz, le isometrie dello spazio iperbolico, che hanno una parametrizzazione più complessa delle semplici matrici ortogonali euclidee.

Nel modello iperboloide, lo spazio iperbolico è rappresentato come il foglio superiore dell'iperboloide $\{x \in \mathbb{R}^{d+1} : x^T \cdot J \cdot x = -1, x_0 > 0\}$ dove $J = \text{diag}(-1, 1, \dots, 1)$ è la matrice di Minkowski. Le isometrie sono trasformazioni lineari Λ che preservano la forma quadratica di Minkowski:

$$\Lambda^T \cdot J \cdot \Lambda = J$$

Queste matrici formano il gruppo di Lorentz $O(1, d)$. Un sottoinsieme importante è costituito dalle rotazioni iperboliche (boosts) lungo una particolare direzione, che possono essere composte per ottenere trasformazioni generali.

L'allineamento di embedding iperbolici può essere formulato come problema di ottimizzazione su questa varietà di Lorentz. Dati insiemi di embedding iperbolici $\{x_i\}$ e $\{y_i\}$ per la lingua sorgente e target, si cerca $\Lambda \in O(1, d)$ che minimizza:

$$\min_{\Lambda} \sum_i d_{\mathbb{H}}(\Lambda \cdot x_i, y_i)^2$$

dove $d_{\mathbb{H}}$ denota la distanza iperbolica. Questo problema è più difficile del corrispondente problema di Procrustes euclideo e richiede algoritmi di ottimizzazione riemanniana specifici.

Un approccio alternativo consiste nel costruire una geodetica tra gli spazi iperbolici sorgente e target. Una geodetica nello spazio delle metriche riemanniane è una curva di deformazione continua che interpola tra due geometrie. Per gli spazi iperbolici, questo può

essere realizzato attraverso il Teichmüller space, uno spazio parametrico di strutture iperboliche su una varietà topologica fissata.

Proiezioni Conformi e Trasformazioni di Möbius

Le trasformazioni conformi sono quelle che preservano gli angoli ma non necessariamente le distanze. Nel contesto degli spazi iperboliche, le trasformazioni di Möbius costituiscono il gruppo delle isometrie conformi del disco di Poincaré.

Una trasformazione di Möbius generale in dimensione d può essere espressa in forma matriciale. Per $d=2$ (il disco complesso di Poincaré), le trasformazioni di Möbius hanno la forma:

$$z \rightarrow (az + b) / (cz + d)$$

dove a, b, c, d sono numeri complessi soddisfacenti ad $-bc = 1$. Questa rappresentazione si generalizza a dimensioni superiori utilizzando algebre di Clifford.

Le trasformazioni di Möbius possiedono proprietà geometriche notevoli: mappano cerchi e linee in cerchi e linee (nel piano esteso), preservano il rapporto anarmonico di quattro punti, e sono completamente determinate dalle immagini di tre punti distinti.

Nel contesto dell'allineamento cross-linguistico, le trasformazioni di Möbius offrono flessibilità aggiuntiva rispetto alle pure isometrie, permettendo di modellare distorsioni locali nella corrispondenza semantica tra lingue mantenendo la struttura conforme. Questo può essere particolarmente utile quando le lingue differiscono nella granularità della distinzione lessicale in diverse regioni semantiche.

Geometria Prodotto e Spazi Misti

Per modellare simultaneamente strutture gerarchiche e non gerarchiche, è possibile utilizzare spazi prodotto che combinano geometrie euclidee, sferiche e iperboliche. Uno spazio prodotto $M = \mathbb{H}^{d_1} \times \mathbb{S}^{d_2} \times \mathbb{R}^{d_3}$ combina una componente iperbolica di dimensione d_1 , una sferica di dimensione d_2 , e una euclidea di dimensione d_3 .

Il tensore metrico dello spazio prodotto è la somma diretta dei tensori metrici delle componenti:

$$g = g_H \oplus g_S \oplus g_R$$

La distanza tra due punti $x = (x_H, x_S, x_R)$ e $y = (y_H, y_S, y_R)$ è data da:

$$d(x,y)^2 = d_H(x_H, y_H)^2 + d_S(x_S, y_S)^2 + d_R(x_R, y_R)^2$$

dove d_H, d_S, d_R sono le distanze nelle rispettive geometrie. La componente sferica può catturare relazioni cicliche o periodiche (come i mesi dell'anno o le ore del giorno), la componente iperbolica modella gerarchie, e la componente euclidea cattura relazioni lineari o distribuite.

Gu e colleghi hanno proposto embedding in spazi prodotto per knowledge graphs, dimostrando miglioramenti significativi nella capacità di catturare simultaneamente relazioni di tipo diverso. L'estensione al contesto multilingue richiede l'allineamento di spazi prodotto, decomponendo il problema in allineamenti separati per ciascuna componente geometrica.

Divergenze Tipologiche e Distanza Linguistica

Le lingue del mondo esibiscono una straordinaria diversità tipologica lungo molteplici dimensioni: ordine dei costituenti (SVO, SOV, VSO, etc.), morfologia (isolante, agglutinante, fusionale, polisintetica), sistema di casi, allineamento sintattico (nominativo-accusativo, ergativo-assolutivo), marcatura di testa vs. marcatura di dipendente, e molte altre. Queste differenze tipologiche si riflettono nella geometria degli spazi di embedding.

Una ipotesi centrale è che lingue tipologicamente simili esibiscano spazi di embedding più facilmente allineabili rispetto a lingue tipologicamente distanti. Questa ipotesi può essere verificata misurando la qualità dell'allineamento (attraverso metriche come P@1 nella traduzione lessicale) in funzione della distanza tipologica.

La distanza tipologica può essere quantificata utilizzando database come WALS (World Atlas of Language Structures), che cataloga centinaia di features tipologiche binarie e categoriche per migliaia di lingue. Date le features tipologiche F_1 e F_2 di due lingue, la distanza di Hamming:

$$d_{\text{typ}}(F_1, F_2) = \sum_i \mathbb{1}(F_1^i \neq F_2^i) / |\{i : F_1^i \text{ e } F_2^i \text{ definiti}\}|$$

fornisce una misura di dissimilarità tipologica normalizzata.

Studi empirici hanno rivelato correlazioni moderate ma significative tra distanza tipologica e difficoltà di allineamento. Lingue all'interno della stessa famiglia (indoeuropee, sinotibetane, austronesiane) tendono ad allinearsi meglio tra loro rispetto a lingue di famiglie diverse, anche quando controllando per la disponibilità di risorse parallele.

Geometria dell'Ordine delle Parole

L'ordine dei costituenti rappresenta una delle dimensioni tipologiche più salienti. Lingue con ordine SVO (soggetto-verbo-oggetto) come l'inglese organizzano le dipendenze sintattiche in modo diverso rispetto a lingue SOV come il giapponese, turco o coreano. Questa differenza si manifesta nella struttura sequenziale delle frasi e, di conseguenza, nella geometria delle rappresentazioni.

Per modellare l'ordine delle parole in modo geometrico, si possono utilizzare rappresentazioni tensoriali delle strutture sintattiche. Un tensore di ordine 3, T_{ijk} , può codificare le probabilità di transizione tra configurazioni sintattiche: l'elemento T_{ijk} rappresenta la probabilità che la configurazione k segua le configurazioni i e j .

La distanza tra due lingue basata sull'ordine delle parole può essere definita come la divergenza di Kullback-Leibler tra le distribuzioni tensoriali:

$$D_{\text{KL}}(T_1 \parallel T_2) = \sum_{i,j,k} T_{1,ijk} \cdot \log(T_{1,ijk} / T_{2,ijk})$$

Valori elevati di questa divergenza indicano differenze strutturali profonde nella organizzazione sintattica.

Un approccio alternativo consiste nell'apprendere trasformazioni che riordinano le rappresentazioni sequenziali. Data una rappresentazione sequenziale di una frase nella lingua sorgente $[h_1, h_2, \dots, h_n]$ (dove h_i sono hidden states di una rete neurale), una trasformazione di riordinamento applica una permutazione π :

$$[h_1, h_2, \dots, h_n] \rightarrow [h_{\pi(1)}, h_{\pi(2)}, \dots, h_{\pi(n)}]$$

Naturalmente, la permutazione ottimale dipende dal contenuto specifico della frase e dalle strutture sintattiche. Quindi, π può essere parametrizzato come funzione del contenuto, appreso mediante reti neurali che predicono scoring per diverse permutazioni, ottimizzate con tecniche come il reinforcement learning o il Gumbel-Softmax per permutazioni differenziabili.

Marcatore Morfologica e Dimensionalità Semantica

Le lingue differiscono marcatamente nella ricchezza del loro sistema morfologico. Lingue morfologicamente ricche come il finlandese, l'ungherese o il turco possono esprimere attraverso suffissi e prefissi informazioni che in lingue analitiche come il cinese o l'inglese richiedono parole separate o costruzioni sintattiche.

Questa differenza ha implicazioni geometriche dirette. In una lingua morfologicamente ricca, una singola radice lessicale può generare centinaia di forme flesse distinte. Se ciascuna forma viene trattata come token separato, lo spazio lessicale diventa vastissimo e frammentato. Alternativamente, si possono apprendere rappresentazioni composizionali dove la forma flessa è ottenuta componendo la radice con marcatori morfologici.

Formalmente, data una radice r e un insieme di marcatori morfologici $\{m_1, m_2, \dots, m_n\}$, la rappresentazione della forma flessa può essere:

$$v_{\text{flessa}} = f(v_r, v_{m_1}, v_{m_2}, \dots, v_{m_n})$$

dove f è una funzione compositiva, ad esempio somma vettoriale, concatenazione seguita da trasformazione non lineare, o prodotti tensoriali.

L'allineamento cross-linguistico tra una lingua morfologicamente ricca e una analitica richiede quindi di gestire questa asimmetria. Una strategia consiste nel segmentare le parole flesse in morfemi prima dell'embedding, ottenendo granularità comparabile. Un'altra consiste nell'apprendere trasformazioni che mappano forme flesse a parafrasazioni analitiche nello spazio target.

Polisemia Cross-Linguistica e Divergenze Lessicali

Le lingue segmentano lo spazio concettuale in modi idiosincratici. Il classico esempio è l'assenza di un singolo equivalente in molte lingue per la parola inglese "sibling" (fratello o sorella senza specificare il genere), che richiede distinzioni di genere in italiano (fratello/sorella), spagnolo (hermano/hermana), etc. Inversamente, molte lingue possiedono

distinzioni lessicalizzate assenti in inglese, come le molteplici parole per "neve" in alcune lingue eschimesi (sebbene l'estensione di questo fenomeno sia spesso esagerata).

Queste divergenze lessicali creano sfide nell'allineamento cross-linguistico. Un singolo concetto in una lingua può mappare a multipli concetti distinti in un'altra, e viceversa. Geometricamente, questo si manifesta come relazioni uno-a-molti o molti-a-molti tra regioni degli spazi di embedding.

Per modellare queste divergenze, si possono utilizzare distribuzioni di probabilità anziché mapping deterministici. Data una parola w nella lingua sorgente, la sua traduzione nella lingua target è una distribuzione:

$$p(t | w) = \text{softmax}(\text{sim}(T(v_w), v_t) / \tau)$$

dove T è la trasformazione cross-linguistica, $\text{sim}(\cdot, \cdot)$ è una misura di similarità (tipicamente prodotto scalare o coseno), v_w e v_t sono gli embedding, e τ è un parametro di temperatura che controlla la concentrazione della distribuzione.

Temperature basse ($\tau \rightarrow 0$) producono distribuzioni piccate concentrate sul vicino più prossimo, mentre temperature elevate distribuiscono la massa di probabilità su molteplici candidati. La scelta ottimale di τ può dipendere dal livello di ambiguità intrinseco nella coppia di lingue.

Proiezioni su Sottospazi Semantici Condivisi

Anziché cercare una trasformazione globale che allinei interamente due spazi di embedding, un approccio alternativo identifica sottospazi semantici condivisi che sono universali tra le lingue, e proietta gli embedding su questi sottospazi.

L'intuizione è che certe categorie semantiche (entità fisiche concrete, numeri, colori) siano più universali e quindi più facilmente allineabili rispetto a categorie culturalmente specifiche (concetti astratti, terminologia giuridica, espressioni idiomatiche). Identificare i sottospazi universali permette di concentrare l'allineamento dove è più affidabile.

Tecnicamente, si può fattorizzare lo spazio di embedding in componenti condivise e specifiche per lingua. Per una lingua l , l'embedding di una parola viene decomposto:

$$v_l = v_{\text{shared}} + v_{\text{specific},l}$$

dove $v_{\text{shared}} \in \mathbb{R}^{d_s}$ appartiene a un sottospazio condiviso di dimensione d_s , e $v_{\text{specific},l} \in \mathbb{R}^{d_p}$ a un sottospazio specifico di dimensione d_p , con $d_s + d_p = d$.

Questa decomposizione può essere appresa mediante tecniche di fattorizzazione matriciale o autoencoders variazionali con vincoli di separazione delle componenti. L'allineamento cross-linguistico viene quindi applicato solo alle componenti condivise, mentre le componenti specifiche catturano idiosincrasie linguistiche non trasferibili.

Distanza Geodetica Cross-Linguistica

Su una varietà riemanniana, la distanza intrinseca tra due punti è definita come la lunghezza della geodetica che li connette. Per varietà immerse in uno spazio euclideo ambiente, la distanza geodetica può essere significativamente diversa dalla distanza euclidea ereditata.

Nel contesto cross-linguistico, si può considerare la varietà prodotto $M = M_s \times M_t$ dove M_s e M_t sono le varietà di embedding delle lingue sorgente e target. Una trasformazione cross-linguistica corrisponde a una sezione di questo prodotto, ovvero una mappa $\sigma: M_s \rightarrow M_t$ tale che ogni punto $x \in M_s$ viene mappato a $\sigma(x) \in M_t$.

La qualità di questa sezione può essere valutata misurando la sua "energia" o "tensione", definita come l'integrale della norma del differenziale:

$$E(\sigma) = \int_{M_s} \|\mathrm{d}\sigma\|^2 \mathrm{dVol}$$

dove $\mathrm{d}\sigma$ è il differenziale di σ (una mappa tra spazi tangenti) e $\|\cdot\|$ è la norma indotta dalle metriche riemanniane. Sezioni di bassa energia corrispondono a trasformazioni "lisce" che non distorcono eccessivamente la geometria.

Minimizzare l'energia porta alle cosiddette mappe armoniche, una generalizzazione delle funzioni armoniche (soluzioni dell'equazione di Laplace) a mappe tra varietà. Le mappe armoniche soddisfano un'equazione differenziale parziale non lineare ellittica che può essere risolta numericamente mediante metodi agli elementi finiti o tecniche variazionali discrete.

Allineamento Basato su Grafi di Conoscenza

Un approccio complementare all'allineamento cross-linguistico sfrutta grafi di conoscenza multilingue come Babelnet, Conceptnet o Wikidata. Questi grafi contengono nodi rappresentanti concetti o entità, e archi rappresentanti relazioni semantiche (iponimia, meronimia, etc.) etichettati con lessicalizzazioni in molteplici lingue.

La struttura del grafo fornisce vincoli geometrici addizionali per l'allineamento. Se due parole in lingue diverse sono etichette dello stesso nodo concettuale, i loro embedding dovrebbero essere allineati. Inoltre, relazioni strutturali nel grafo (due nodi connessi da un cammino) dovrebbero essere preservate attraverso la trasformazione.

Formalmente, sia $G = (V, E)$ un grafo di conoscenza multilingue con nodi V e archi E . Per ciascun nodo $v \in V$, sia $L(v)$ l'insieme di lessicalizzazioni di v nelle diverse lingue. L'obiettivo è apprendere embedding e_v per ciascun nodo v che siano consistenti con le relazioni del grafo e allineati con gli embedding lessicali.

Una funzione obiettivo combinata potrebbe essere:

$$L = \sum_{(v,r,u) \in E} d(e_v + r, e_u)^2 + \sum_v \sum_{l \in L(v)} d(e_v, \text{emb}_l(l))^2$$

Il primo termine impone che la relazione r trasli e_v vicino a e_u (come nei modelli TransE per knowledge graph embeddings), mentre il secondo termine allinea gli embedding dei nodi con gli embedding lessicali delle loro etichette.

Analisi delle Componenti Principali Cross-Linguistiche

L'analisi delle componenti principali (PCA) può essere estesa al setting multilingue per identificare direzioni di varianza condivise tra le lingue. Data una collezione di matrici di embedding $\{X_1, X_2, \dots, X_L\}$ per L lingue, si può concatenare orizzontalmente (dopo opportuno allineamento preliminare) ottenendo una matrice combinata X_{combined} .

Applicando PCA a X_{combined} , le componenti principali risultanti catturano direzioni di massima varianza che sono consistenti attraverso le lingue. Le prime componenti spesso corrispondono a dimensioni semantiche universali (concretezza vs. astrattezza, animatezza, positività vs. negatività), mentre le componenti successive possono catturare variazioni più specifiche.

Una analisi più raffinata utilizza la Consensus PCA, che cerca componenti principali che massimizzano simultaneamente la varianza spiegata in ciascuna delle L lingue. Formalmente, si cerca la matrice di proiezione W che massimizza:

$$\max_W \sum_L \text{tr}(W^T X_L X_L^T W)$$

soggetto a $W^T W = I$. Questa è una generalizzazione multi-vista della PCA standard che può essere risolta mediante tecniche di autodecomposizione generalizzata.

Invarianti Topologici e Caratterizzazione Geometrica

La topologia algebrica fornisce strumenti per caratterizzare proprietà globali degli spazi di embedding attraverso invarianti topologici come gruppi di omologia e numeri di Betti. I gruppi di omologia $H_k(M)$ per $k = 0, 1, 2, \dots$ catturano i "buchi" di dimensione k nella varietà M : H_0 conta le componenti connesse, H_1 i cicli unidimensionali (come anelli), H_2 le cavità bidimensionali (come sfere vuote), etc.

I numeri di Betti $\beta_k = \text{rank}(H_k)$ forniscono una firma topologica della varietà. Per gli spazi di embedding linguistici, analisi di omologia persistente hanno rivelato strutture topologiche non triviali. L'omologia persistente traccia l'evoluzione della topologia al variare di un parametro di scala, tipicamente il raggio di filtrazione in una costruzione di complesso simpliciale. Dato un insieme di punti (embedding) $P \subset \mathbb{R}^d$ e un raggio r , si costruisce il complesso di Čech $C_r(P)$ i cui simplessi sono sottoinsiemi di punti tali che le palle di raggio r centrate su di essi hanno intersezione non vuota. Al crescere di r , nuovi simplessi vengono aggiunti e la topologia evolve. L'omologia persistente traccia quando features topologiche (componenti connesse, cicli, cavità) appaiono e scompaiono, producendo un diagramma di persistenza che codifica la struttura topologica multi-scala.

Confrontando i diagrammi di persistenza di spazi di embedding di lingue diverse mediante metriche come la distanza di bottleneck o la distanza di Wasserstein tra diagrammi, si possono quantificare similarità topologiche cross-linguistiche. Lingue tipologicamente simili potrebbero esibire diagrammi di persistenza simili, riflettendo strutture semantiche condivise.